Ben Gowaski

EECE5645 Parallel Processing

Assignment 1: Text Analyzer

**Question 0:** Go to the directory that contains TextAnalyzer.py and run the following from the command prompt:

python TextAnalyzer.py --help

What does this print?

A screenshot of a cell phone

Description automatically generatedWhat portion of the code causes this to be printed?

-The argparse section at the top of main statement allows this to be printed as so. <https://docs.python.org/2/library/argparse.html>

**Question 1:**

**1(a)**

**A close up of text on a black background

Description automatically generated**

**1(b):**

python TextAnalyzer.py TF \ masc\_500k\_texts/written/fiction/hotel-california.txt \ hotel-california.tf

**output from first 5 lines of generated text file:**

[gowaski.b@c0081 Assignment1]$ vi hotel-california.tf/part-00000

('sound', 8)

('is', 27)

('vibration', 2)

('as', 27)

('mechanical', 1)

**Question 2: Find the words with highest frequency.**

**2(a):**

**A screenshot of a cell phone

Description automatically generated**

**2(b): Top 20 most frequent words in doc**

[gowaski.b@c0081 Assignment1]$ vi hotel-california.top

(the, 294)

(i, 192)

(to, 145)

(of, 142)

(and, 137)

(a, 136)

(was, 102)

(in, 76)

(it, 68)

(he, 58)

(my, 57)

(that, 53)

(on, 51)

(we, 50)

(you, 48)

(all, 48)

(had, 47)

(me, 44)

(said, 42)

(were, 41)

**Question 3:** Find the 20 most frequent words in the entire corpus

[gowaski.b@c0081 Assignment1]$ vi all.top

(the, 26054)

(to, 13457)

(and, 12370)

(of, 11981)

(a, 10353)

(in, 8299)

(that, 6464)

(i, 6433)

(is, 5318)

(you, 4717)

(for, 4227)

(it, 4012)

(on, 3372)

(with, 3136)

(was, 2994)

(as, 2859)

(this, 2759)

(be, 2520)

(we, 2447)

(have, 2319)

**Question 4: Calculate IDF**

**4(a):**

**A screenshot of a cell phone

Description automatically generated**

**4(b):**

[gowaski.b@c0155 Assignment1]$ vi anc.idf/part-00000

('ellen', 4.174387269895637)

('kind', 1.761454119732726)

('of', 0.08561375272299172)

('do', 0.8303483020734302)

('think', 1.341173925839421)

**Question 5:**

**5(a)**

**A close up of text on a black background

Description automatically generated**

**5(b) 20 TFIDF scores of file hotel-california.txt**

[gowaski.b@c0155 Assignment1]$ vi hotel-california.top1

(adrienne, 179.2819849911674)

(ship, 117.23397267614868)

(i, 102.22554922933487)

(zheng, 102.21822345956723)

(ray, 87.13417653379183)

(sarah, 83.48774539791273)

(kishori, 68.54899426132872)

(was, 66.5999575283543)

(he, 62.83399935916463)

(she, 62.377099674377554)

(my, 60.89349553284851)

(tiffany, 58.41041340546698)

(captain, 52.263075670888924)

(her, 51.06110233379411)

(jefferson, 48.242839080820545)

(glass, 47.903370253258664)

(said, 47.788986076498425)

(had, 44.28651225903854)

(looked, 42.39411887663042)

(me, 41.45971360420629)

**Question 6: Cosine Similarity**

**6(a)**

**A screenshot of a cell phone

Description automatically generated**

python TextAnalyzer.py TF masc\_500k\_texts/written/fiction fiction.tf

python TextAnalyzer.py IDF fiction.tf/ fiction.idf

python TextAnalyzer.py TFIDF fiction.tf fiction.tfidf --idfvalues anc.idf

python TextAnalyzer.py TF masc\_500k\_texts/written/spam spam.tf

python TextAnalyzer.py IDF spam.tf/ spam.idf

python TextAnalyzer.py TFIDF spam.tf spam.tfidf --idfvalues anc.idf

python TextAnalyzer.py SIM fiction.tfidf fiction-spam.sim --other spam.tfidf

[gowaski.b@c0069 Assignment1]$ vi fiction-spam.sim

fiction.tfidf, spam.tfidf, 0.3233081294319646

fiction.tfidf, fiction.tfidf/, 1.0000000000000009

|  |  |  |
| --- | --- | --- |
|  | fiction | spam |
| fiction | 1.0000000000000009 | 0.3233081294319646 |
| spam | 0.3233081294319646 | 1.0000000000000009 |